

The person charging this material is responsible for its return to the library from which it was withdrawn on or before the **Latest Date** stamped below.

Theft, mutilation, and underlining of books are reasons for disciplinary action and may result in dismissal from the University.

To renew call Telephone Center, 333-8400

UNIVERSITY OF ILLINOIS LIBRARY AT URBANA-CHAMPAIGN

DEC 6 1991



Digitized by the Internet Archive
in 2013

<http://archive.org/details/predictinggradua925nels>

510.84
Ill
UIUCDCS-R-78-925
no. 925
copy 2

Math

UILU-ENG 78 1717

8

PREDICTING GRADUATE SCHOOL SUCCESS

by

Richard H. Nelson

June 1978



DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · URBANA, ILLINOIS

The Library of the
JUL 1978
University of Illinois

UIUCDCS-R-78-925

PREDICTING GRADUATE SCHOOL SUCCESS

by

Richard H. Nelson

Spring 1978

Submitted in partial fulfillment of the requirements for the degree of
Master of Computer Science in the Graduate College of the University of
Illinois at Urbana-Champaign, 1978

TABLE OF CONTENTS

	Page
INTRODUCTION.....	1
STATISTICAL OVERVIEW.....	3
METHOD.....	7
RESULTS AND DISCUSSION.....	11
SUMMARY AND RECOMMENDATIONS.....	24
APPENDICES.....	29

INTRODUCTION

The purpose of this research was to try to identify the variables affecting "success" in graduate school in computer science. The motivation behind the study was the desire to obtain an objective criterion for predicting an individual's chances of success. Such a criterion would be a valuable tool in facilitating the process of graduate admissions. Evaluations would be standard, result in better prediction, and make selections simpler and faster.

The subjects for the study were entering graduate students to the Department of Computer Science at the University of Illinois from spring 1974 to fall 1975. In all there were 170 entering graduate students but adequate data existed for only 73.¹ These 73 form the sample under investigation.

Before attempting to predict success in graduate school, it was first necessary to form an operational definition of what constitutes graduate school success. There are several measurable achievement criteria, two of which are first semester grade point average (GPA) and cumulative graduate GPA. A third available measure was the number of times the Department's general exams had to be taken by each student. Data were collected on each of these measures and multiple regression analyses were performed.

¹ Although the research was exploratory, it was felt that GRE scores (particularly the GRE quantitative score) and undergraduate GPA were important enough variables to warrant the exclusion of any case lacking data on either one. As it turned out, GPA could usually be estimated from the transcript, and thus the vast majority of excluded cases were lacking, at the very least, GRE scores.

But these were not the main foci of the study. Although each of these variables was easily measurable and achievement-oriented, it was not felt that any one of them truly captured what is meant by a "successful" graduate school career. A student obtaining a GPA of 5.00 the first semester but then leaving graduate school cannot be considered successful. Similarly, a student who passes each general exam on the first attempt but drops out of school after that semester would not ordinarily be considered successful. Actually, the achievement of obtaining a graduate degree is ultimately the best measure of graduate school success, and a student who earns a Ph.D. is ordinarily thought to be more successful than one who earns a master's degree. This definition of success is probably one with which most department heads and admissions' committees could be comfortable. This degree criterion is thus the major dependent variable of this research. Since several years had elapsed since the students originally entered graduate school, it was possible to divide the sample into three groups: one consisting of those students who dropped out of the department before receiving any graduate degree; a second comprising those who received masters' degrees and then departed;² and a third group consisting of those who passed the department's qualifying exam, a necessary precondition for study leading to the Ph.D. All members of the second group either did not take the qualifying exam or failed it, and thus the achievement of

²No distinction was made between the different master's programs since, of the 28 students in the master's group, only one received an M.C.S. and one an M.S.T.C.S.

passing the qualifying exam is taken as an indication of eventual completion of the Ph.D.³

The statistical technique of discriminant analysis was used to determine the dimensions along which the three groups differ. The results of this analysis form the first portion of the following discussion. The strength of the relationship obtained is then discussed by viewing discriminant analysis as a special case of canonical correlation. Finally, the results of the discriminant analysis -- the predictor variables and the goodness of the prediction -- are compared to the results of the multiple regressions mentioned above. This is done to ascertain whether the two procedures do indeed yield different results and thus, to what extent the degree criterion proposed differs from measures such as GPA.

But before proceeding to the details of the analysis, a brief discussion is offered of the major statistical techniques employed. A more extended treatment can be found in Tatsuo (1971).

STATISTICAL OVERVIEW:

Multiple regression is the statistical technique whose goal is to form a linear combination of the independent variables,

$(X_1, \dots, X_p)$

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \dots + b_pX_p$$

³The original research design called for studying a class of students who had all ended their graduate school careers. The poor quality of the data available from administrative sources and the fact that the department's computerized records only date back to 1974 made this infeasible.

such that the correlation ($r_{y\hat{y}}$) between this linear combination and the dependent variable (Y) is maximal. An equivalent formulation in terms of least squares gives the goal of multiple regression as minimizing the sum of squared discrepancies, i.e., finding a linear combination of the X's

such that $\sum_{i=1}^N (Y_i - \hat{Y}_i)^2 = \text{minimum}$, where N = number of observations.

The solution is obtained by taking the partial derivative of the expression

$$\sum_{i=1}^N (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^N (Y_i - b_0 - b_1 X_{i1} - \dots - b_p X_{ip})^2 \text{ with respect to each}$$

of the variables $b_0, b_1, \dots, b_p$, setting each derivative to zero, and

solving the resulting set of linear equations for $b_0, \dots, b_p$. The

equation can then be used to predict Y (say GPA) for incoming students.

Discriminant analysis is the appropriate statistical technique when the sample is divided into two or more clearly-defined groups, and the object of the research is to determine how the groups differ with respect to the p independent variables in the system. The technique is conceptually similar to multiple regression, but the dependent variable is now categorical rather than continuous. The goal of discriminant analysis is the formation of a linear combination of the independent variables which maximally differentiates between the K groups. The criterion of differentiation is the ratio of the between-groups to the within-groups sums of squares on the new variable Y (the linear combination of the independent variables). The between-groups sums of squares

$$(SS_b(Y)) \text{ equals } \sum_{k=1}^K n_k (\bar{Y}_k - \bar{Y})^2, \text{ where } n_k \text{ is the number of subjects}$$

in group k , $\bar{Y}_k$ is the mean of Y for group k , and $\bar{Y}$ is the grand mean of all the subjects on Y . The within-groups sums of squares ($SS_w(Y)$) can

$$\text{be expressed as } SS_1(Y) + SS_2(Y) + \dots + SS_K(Y) = \sum_{k=1}^K \sum_{i=1}^{n_k} (Y_{ik} - \bar{Y}_k)^2,$$

where Y_{ik} is the score of the i th individual in the k th group on variable Y .

Solution of the problem is readily obtained upon solving the eigenvector-eigenvalue problem $(\underline{W}^{-1} \underline{B} - \lambda \underline{I}) \underline{v} = \underline{0}$, where

$\underline{W}$ = the within-groups sums of squares and cross products (SSCP) matrix

$\underline{B}$ = the between-groups SSCP matrix

$\underline{v}$ = the vector of combining weights for the p independent variables

$$\lambda = SS_b(Y) / SS_w(Y).$$

In general, solution of the eigenvalue problem results in the identification of $\min(p, K-1) = r$ eigenvalues and associated eigenvectors. So in general, discriminant analysis produces r discriminant functions such that the first function is that linear combination of the independent variables which yields the largest discriminant criterion λ_1 . The second function corresponds to λ_2 , the largest discriminant criterion of any linear combination of the independent variables which is uncorrelated with the first linear combination. Similarly, the r th discriminant function yields λ_r , the largest discriminant criterion of any linear combination of the variables which is uncorrelated with the previous $r-1$ discriminant functions. Variables which make large contributions to between-group discrimination

are those which would be most useful in selecting prospective graduate students.

Canonical correlational analysis is applicable when dealing with two sets of variables (e.g., q dependent as well as p independent variables). The goal is to form linear combinations of each of the two sets such that the correlation between the linear combinations is a maximum. Determining the sets of coefficients is accomplished, after appropriate manipulations, by solving the eigen equation

$$(\underline{S}_{xx}^{-1} \underline{S}_{xy} \underline{S}_{yy}^{-1} \underline{S}_{yx} - \mu^2 \underline{I}) \underline{u} = \underline{0}$$

where $\underline{S}_{xx}$ = the SSCP matrix for the X set of variables

$\underline{S}_{yy}$ = the SSCP matrix for the Y set of variables

$\underline{S}_{xy}$ = the p x q cross products matrix between the X and Y sets of variables

$\underline{S}_{yx}$ = the q x p cross products matrix between the Y and X sets of variables

μ = the correlation between the linear combinations

$\underline{u}$ = the vector of coefficients for the X set.

In general, solution of the eigen equation yields $\min(p, q) = r$ pairs of u and v vectors. Each pair of linear combinations is called a canonical variate. The interpretation is that the first canonical variate is that pair of linear combinations of the two sets of variables which is maximally correlated. The second canonical variate produces the maximum correlation between linear combinations which are uncorrelated with the first variate. And the rth canonical variate yields the maximum

correlation between any pair of linear combinations uncorrelated with the first $r-1$ canonical variates.

METHOD:

The independent and dependent variables for which data were collected are listed in Table 1. Where appropriate, an explanation of the variable, scaling technique, or coding method is included. The actual raw data values are given in Appendix 1.

With a total sample size of only 73, it was necessary to find some way to limit the number of independent variables included in the discriminant analysis. Although inclusion of all ten variables would maximize the discriminating power in the given sample, the coefficients of the resulting discriminant functions would be highly related to the sampling errors present since the greater the number of variables relative to sample size, the greater the effect of sampling error and the higher the discrepancies between the computed coefficients and those which would be obtained from a different sample. Since the goal of this study was not to maximize the variability explained for the given sample, but to obtain reasonable prediction for future samples, it was thought that a moderate number of independent variables, relative to the sample size, should be selected. Due to the exploratory nature of the research, there were no a priori grounds upon which to select a subset of the variables. The procedure followed was to first run a stepwise discriminant analysis which utilized a forward selection technique, i.e., variables were included in the analysis one-by-one

Table 1

<u>Variable</u> ³	<u>Explanation</u>
Independent:	
undergraduate GPA	Scale of 1.00 to 5.00
GRE verbal percentage	
GRE quantitative percentage	
average score over three letters of reference	Scale of 0 to 4 with 0 = poor and 4 = superior
Astin selectivity index ⁴ of undergraduate college attended	Scale of 1 to 7 -- the index is based on the average ACT and SAT scores of entering freshman at the college -- it thus measures the general ability level of the students and the selectivity level of the college (see Appendix 2 for sample Astin scores)
undergraduate major	0 = computer science 1 = CS-related field (e.g. engineering or math) 2 = field unrelated to computer science
highest prior degree obtained	0 = bachelor's 1 = master's
# of C.S. courses taken as an undergraduate	
GPA in C.S. courses	
interaction between verbal GRE score and whether student was foreign	foreign = 0 if no, 1 if yes interaction = foreign * GRE verbal
Dependent:	
1st semester graduate GPA	If student dropped out prior to end of first semester, GPA = 2.00
cumulative graduate GPA	
number of times the general exams were taken	To obtain an M.S., general exams must be passed in four substantive areas. Each exam can be taken a maximum of 3 times. A score of 4 was assigned to an exam which was failed 3 times or which had not been passed when the student left school. The range of scores was thus 4 to 16: 4 if each test was passed the first time, and 16 when each test was never passed.

³ See Appendices 3 and 4 for correlation matrices of independent and independent vs. dependent variables, respectively.

⁴ No Astin rating was available for foreign institutions. To obtain estimated scores, the Astin index was first stepwise-regressed on the other independent variables and this equation was applied to the 16 foreign students. For $\alpha = .05$, there were three significant predictors and a multiple R of .50468. The equation was:

$$\text{Astin} = 2.234 - 1.147 * \text{largest degr} - 1.165 * \text{undergrad GPA} + .0874 * \text{GRE quant}$$

in such a way that the i th variable selected was that variable which added the most to the discriminating power given by the first $(i-1)$ variables. Given the order of variable selection, analyses were then run using the "best" i variables, $3 \leq i \leq 10$, and appropriate statistics were collected which reflected the "goodness" of the discriminating power for that particular number of independent variables. Finally, the aim was to select for further analysis the smallest number of variables while retaining as much predictive power as possible.

From this procedure it was decided to retain the best six variables for inclusion in the following analyses and discussion. There are several justifications for this decision. First, it is seen from Table 2 that when using six variables, the maximum percentage for correctly classifying cases into groups is reached. Second, the obtained canonical correlations of .548 and .431 between group membership and the independent variables are reasonably close to the maximum achieved using all the variables. Finally, although the correlations obtained using four or five variables are relatively close to those for six variables, there are theoretical reasons for retaining all six variables. It is seen that the interaction term between verbal GRE scores and foreign student classification was the sixth variable selected. Exclusion of this variable from the analysis befuddles any interpretation of the effect of GRE verbal scores, since the sample correlation between verbal GRE and foreign student classification was $-.79$. Given that verbal GRE is to be included in the analysis (it was stepwise selected

Table 2

Summary of stepwise discriminant analyses

	Canonical correlations	% of cases correctly classified
with best 3 variables	.479, .291	50.68%
with best 4 variables	.486, .411	60.27%
with best 5 variables	.511, .430	60.27%
with best 6 variables	.548, .431	64.38%
with best 7 variables	.557, .457	63.01%
with best 8 variables	.561, .465	63.01%
with best 9 variables	.565, .480	60.27%
with all 10 variables	.566, .483	64.38%

Order of inclusion:

1. Astin selectivity index
2. letters of reference
3. GRE verbal
4. GRE quantitative
5. undergraduate major
6. verbal GRE-foreign student interaction
7. highest prior degree
8. undergraduate GPA
9. GPA in undergraduate CS courses
10. number of undergraduate CS courses taken

third), its interaction with the foreign classification should be included also. Reducing the effect of sampling errors (as manifested by the percentage of foreign students in the given sample) also argues for inclusion of the interaction term. For all the above reasons, the following results are based on the effects of six independent variables.

RESULTS AND DISCUSSION:

Discriminant Analysis: With $K=3$ groups and $p=6$ variables, two discriminant functions are obtained. Their significance can be seen from the following: Given that $\lambda_1 = .42921$ and $\lambda_2 = .22755$, Bartlett's V is seen to be

$$\begin{aligned} V &= (N-1-(p+K)/2) \sum \ln(1+\lambda_i) \\ &= 24.106 + 13.839 \\ &= 37.945 \end{aligned}$$

which is approximately distributed as a chi-square with $p(K-1) = 12$ degrees of freedom provided $N-1-(p+K)/2$ is large (Tatsuoka, p. 164). Since $\chi^2_{12,.001} = 32.909$ the difference in group centroids is highly significant. To see that both functions are significant it is only necessary to look at $V_2 = V - V_1 = 13.839$ with $(p-1)(K-2) = 5$ degrees of freedom. Since $12.833 = \chi^2_{5,.025} < V_2 < \chi^2_{5,.01} = 15.086$ the second discriminant function is significant at the .025 level.

In attempting to interpret the functions and the corresponding

dimensions along which the groups differ, first consider the plot of cases and group means on each function given in Table 3. It is seen that on function one the means of groups 1 (dropout) and 2 (master's) are nearly identical while that of group 3 (Ph.D.) is considerably less; whereas on function two the groups are spaced at roughly equal intervals with group 2 having the highest mean and group 1 the lowest. To determine the relative contribution of each original variable to each discriminant dimension it is necessary to look at the standardized discriminant weights given in Table 4.

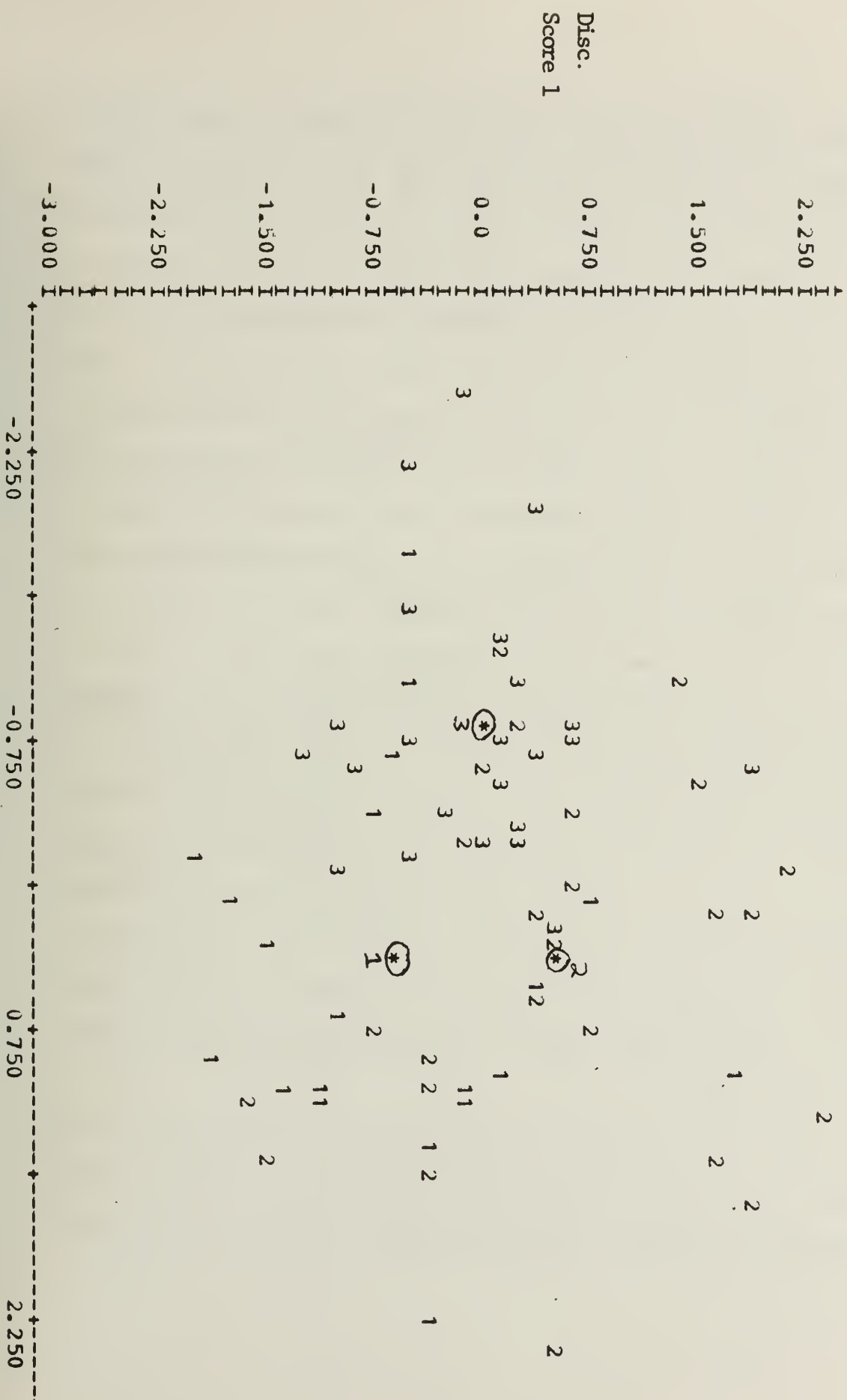
Table 4

Standardized Discriminant Weights		
	Function 1	Function 2
Astin index	-.30133	.99178
letters of recommendation	-.38304	.31091
GRE verbal	.37866	.38702
GRE quantitative	-.38939	-.57703
undergraduate major	.40701	.24461
verbal-foreign interaction	-.39696	-.05164

These coefficients allow for no simple labelling of the two dimensions. On the first dimension, each original variable contributes roughly equally. Thus, a low score on dimension one (e.g. the Ph.D. group mean) is associated with a high Astin index, high letters of recommendation, high GRE quantitative, undergraduate major in the direction of computer science, low GRE verbal, and high GRE verbal-foreign student interaction. The first four characteristics are what one would expect of the Ph.D. group. The low verbal GRE scores and high verbal GRE-foreign interaction indicate a higher proportion of

Table 3

Plot of Discriminant Scores



foreign students in the Ph.D. group than the other two groups. This is seen to be the actual case, as the Ph.D. group is composed of 39% foreign students, the master's group 11%, and the dropout group 23%.

Given the interpretation of the first function, the second function is best viewed as a discriminator between groups one and two. The position of the group three mean between those of groups one and two makes any interpretation of this dimension which includes group three difficult at best. The conceptual simplicity obtained by viewing dimension one as separating the Ph.D. group from the other two and dimension two as separating the master's from the dropout group also argues strongly for adopting this schemata. When this is done the magnitudes of the standardized discriminant weights show that the Astin index is by far the primary variable involved in this dimension, with GRE quantitative being roughly half as important but in the opposite direction, and the rest occupying a background role. Thus someone scoring high on dimension two (e.g. the master's group mean) would tend to have a very high Astin index and a relatively low GRE quantitative score. Theoretically, this runs counter to expectations. It would be expected that the master's group would have a higher mean score on both the Astin index and the GRE quantitative variables than the dropout group.

At this point it may be instructive to look at the structure matrix in an effort to gain further insight into the meaning of the discriminant dimensions and perhaps explain the apparently contrary

result found above. The structure matrix is the $p \times (K-1)$ matrix of correlations between the p original variables and the $K-1$ discriminant functions. The coefficients are given in Table 5.

Table 5
Discriminant analysis structure matrix

	Function 1	Function 2
Astin index	-.4313	.7336
letters of recommendation	-.4969	.2449
GRE verbal	.3056	.4505
GRE quantitative	-.5729	.0583
undergraduate major	.4648	.1765
verbal-foreign interaction	-.3822	-.2398

These correlations corroborate the interpretation given to the first discriminant dimension. Although the numbers have changed slightly, they are still within a factor of two of each other with the same signs, and thus the original variables can still be considered to be of roughly equal importance to the first discriminant dimension.

Utilizing the correlations from the structure matrix fortunately results in a different and more harmonious interpretation of the second discriminant dimension. While the Astin index still rates as the largest contributor to this dimension, the effect of quantitative GRE scores is seen to be eliminated, with the verbal GRE variable taking its place. The phenomenon whereby it is possible for a variable to have a high negative loading on a discriminant function but a low correlation with the function is analagous to the situation in multiple regression

where the standardized regression weight is large in absolute value but the zero-order correlation with the dependent variable is small. Both situations arise from the nature of the intercorrelations of the independent variables. Discriminant function weights (or regression weights) control for the intercorrelations. The correlations from the structure matrix are zero-order correlations, with the resultant possibility of suppression or spuriousness.

Thus the profile of an individual scoring high on the second discriminant function is one of a high Astin index score and a relatively high GRE verbal score. This does not contradict any theoretical expectations, but does confirm what one might guess from a glimpse of the group means listed in Table 6 and a reference back to the percentage of foreign students in the two groups.

Table 6

Group means on independent variables

Independent variables used	Dropout group	Master's group	Ph.D. group	Total
letters of reference	2.7509	2.9107	3.1548	2.9394
undergraduate major	0.7273	0.8214	0.4348	0.6712
GRE verbal	56.9091	69.9643	53.1304	60.7260
GRE quantitative	89.0000	90.0714	96.6956	91.8356
Astin index	4.0673	5.1986	5.3696	4.9115
verbal-foreign interaction	4.1818	1.6429	8.0000	4.4110
Variables not used				
undergraduate GPA	4.3495	4.4075	4.5735	4.4423
highest degree	0.3182	0.0714	0.0870	0.1507
CS course GPA	4.7127	4.6279	4.7126	4.6801
# of CS courses	4.5000	3.7857	5.2174	4.4521

Once again, the effect of the verbal GRE variable is largely due to its interaction with the foreign student classification variable. It is therefore tempting to simplify the interpretation of dimension two even further and claim that the major difference between the master's group and the dropout group is largely one of a difference in the competitiveness of the undergraduate institutions attended by members of the two groups.

Turning from the area of interpretation to that of prediction leads to the final results obtained from the discriminant analysis, the classification functions and predictions.

The classification functions obtained from the statistical software package SPSS (Statistical Package for the Social Sciences) are derived from the pooled within-groups covariance matrix and the centroids for the discriminating variables. There is one function for each group. Scores on the three functions are obtained by multiplying the raw scores on each of the variables by the appropriate coefficient and adding in the constant. A case would be assigned to that group for which its classification function score is the highest. Table 7 lists the coefficients. The results obtained by applying the classification functions to the original sample are given in Table 8.

The table shows that about two-thirds of the original sample was correctly classified using the derived classification functions. This seems to be a reasonable figure, and it becomes even more reasonable when the group breakdowns are taken into account. It would be most harmful to incorrectly classify someone whose actual group membership

Table 7

Classification function coefficients

	Dropout gp.	Master's gp.	Ph.D. group
letters of recommend.	10.28953	11.05553	11.81329
undergraduate major	5.19958	5.62126	4.35394
GRE verbal	-0.00784	0.00732	-0.02119
GRE quantitative	0.97513	0.91136	1.00213
Astin index	-1.43927	-0.55646	-0.64228
verbal-foreign	0.10469	0.10248	0.16021
constant	-56.50551	-58.33578	-66.38499

Table 8

Predicted Group Membership

<u>Actual Group</u>	<u>N</u>	<u>Group 1</u>	<u>Group 2</u>	<u>Group 3</u>
1 - Dropout	22	13 59.1%	5 22.7%	4 18.2%
2 - Master's	28	6 21.4%	16 57.1%	6 21.4%
3 - Ph.D.	23	2 8.7%	3 13.0%	18 78.3%

% of grouped cases correctly classified: 64.38%

turned out to be the Ph.D. group. Recognizing as many potential Ph.D.'s as possible is highly desirable and, for the given sample, this turned out to be exactly the case. The Ph.D. group was by far the most distinguishable, with 18 out of the 23 cases in that group being correctly classified.^{5,6}

Strength of relationship: The next topic to be discussed is the strength of the relationship found in the discriminant analysis. There is no directly comparable statistic in discriminant analysis to the multiple correlation coefficient of regression analysis, but treating discriminant analysis as a special case of canonical correlation allows the computation of various measures of strength of relationship.

When the K groups of the discriminant analysis are coded as K-1 dummy variables (i.e., 1 for group membership, 0 for non-members), a canonical correlation analysis can be performed, resulting in the identification of K-1 canonical variate pairs and K-1 canonical correlations. The two canonical correlations obtained in the present case were .548 and .431. However, although the correlations can be considered to be indicators of the strength of the relationship, the

⁵This is not surprising, since the discriminant function corresponding to the largest discriminant criterion was interpreted as being a discriminator between the Ph.D. group and the other two groups.

⁶There are no obvious generalizations to make regarding the five incorrectly classified Ph.D. students, but individual data on each are given in Appendix 7.

interpretation in terms of explained variance is not obvious. A squared canonical correlation represents the proportion of variance of a linear combination of one set of variables that is predicted by a linear combination of the other set of variables. What is desired is not a measure of the variance shared by linear combinations of the two sets of variables, but rather a measure of the shared variance of the two sets. Stewart and Love's (1968) redundancy coefficients provide such a measure. These coefficients are asymmetrical measures of the strength of the relationship between two sets of variables. There is one coefficient for the redundancy of set Y given set X, and another for the redundancy of set X given set Y. In the present case, there is a need for only one of the coefficients, since the sets of variables are clearly delineated into dependent and independent categories. Letting Y = the set of 2 dummy group variables and X = the set of 6 independent variables, the coefficient of interest is $\overline{R_{Y \cdot X}^2}$, which is the redundancy of set Y given set X or, equivalently, the proportion of variance in Y predictable from X. Computing $\overline{R_{Y \cdot X}^2}$ from the formula,

$$\overline{R_{Y \cdot X}^2} = \sum_{i=1}^q \mu_i^2 \left(\sum_{j=1}^q a_{yji}^2 \right) / q$$

where μ_i^2 = the *i*th squared canonical correlation

a_{yji}^2 = the squared correlation between Y and the *i*th canonical variate of the Y set

$q = 2$, the number of variables in Y

yields the result that $\overline{R^2_{y \cdot x}} = .2562$. Thus, approximately 25% of the variance in the group classification is predictable from the independent variables. The corresponding $\overline{R_{y \cdot x}}$ is .506.

Multiple Regression: The final section briefly presents the results of the three multiple regression analyses where first semester graduate GPA, cumulative graduate GPA, and general exam cumulative total were the dependent variables. A contrast is then offered with the discriminant results.

The regressions were also done in a stepwise fashion, so it is possible to get an idea of the importance of each of the independent variables in the separate predictions. Table 9 lists the multiple correlations obtained at each step in the selection process for each of the dependent variables. (Appendix 6 gives the corresponding regression weights.) Several conclusions can be reached upon examination of Table 9. The first is that there is little to be gained by including more than two or three predictors for each dependent variable. The multiple correlations level off quite rapidly. Second, the specific nature of each dependent variable can be seen from its predictors. The number of times the general exams were taken is largely a function of an individual's mathematical aptitude and his undergraduate background in computer science. First semester grade point average and cumulative grade point average are virtually indistinguishable, the correlation between them being .92,

Table 9

Regression summaries

Dependent variable = 1st semester graduate GPA

<u>Variable</u>	<u>Multiple R</u>	<u>R</u>	<u>Change in R</u>	<u>simple r</u>
Astin index	.44192	.19530	.19530	.44192
undergraduate GPA	.55133	.30397	.10867	.27711
largest degree	.57205	.32725	.02328	-.01959
verbal-foreign	.57934	.33563	.00838	.07668
# CS courses	.58744	.34508	.00946	.19304
GRE quantitative	.58896	.34687	.00179	.35515
GRE verbal	.58991	.34799	.00112	.09403
undergraduate major	.59062	.34884	.00084	-.16153
CS GPA	.59084	.34909	.00025	.14825
letters of reference	.59089	.34915	.00006	.11604

Dependent variable = cumulative graduate GPA

<u>Variable</u>	<u>Multiple R</u>	<u>R</u>	<u>Change in R</u>	<u>simple r</u>
Astin index	.39665	.15733	.15733	.39665
undergraduate GPA	.53071	.28165	.12432	.30506
verbal-foreign	.54060	.29225	.01060	.06733
undergraduate major	.55209	.30480	.01255	-.22006
largest degree	.55802	.31138	.00658	-.07921
CS GPA	.56284	.31678	.00540	.10366
GRE quantitative	.56343	.31745	.00067	.36888
# CS courses	.56383	.31790	.00045	.17834
GRE verbal	.56406	.31817	.00026	.09745
letters of reference	.56416	.31828	.00011	.16847

Dependent variable = general exam cumulative total

<u>Variable</u>	<u>Multiple R</u>	<u>R</u>	<u>Change in R</u>	<u>simple r</u>
GRE quantitative	.44696	.19977	.19977	-.44696
# CS courses	.56043	.31408	.11431	-.43701
Astin index	.59140	.34975	.03567	-.37221
undergraduate GPA	.62592	.39177	.04202	-.30076
CS GPA	.64095	.41082	.01905	-.01930
verbal-foreign	.65208	.42521	.01439	-.05166
undergraduate major	.65816	.43318	.00797	.33509
largest degree	.65821	.43324	.00006	.11199
GRE verbal	.65821	.43324	.00000	-.05815
letters of reference	.65821	.43324	.00000	-.21739

and both are functions of undergraduate grade point average and the quality of the undergraduate institution. Thus, although grade point average and number of times the general exams were taken are achievement-related variables, the achievements they are related to are rather narrow, specific criteria. This contrasts rather sharply with the notion of treating the end product of graduate school as the dependent variable, as was done in the discriminant analysis. When the classification into Ph.D., master's, and dropout groups was made, the concepts of achievement and success become much more general and all-encompassing. Thus, although the relationship between the group classification scheme and grade point average is apparent, the groups represent more than grade point average; they represent graduate school achievement in a more general yet meaningful sense of the term.

Comparison of the redundancy coefficient given earlier with the multiple correlations obtained in the regression analyses shows that the predictive power is not as great when the dependent variables consist of the discriminant groups. This is explainable in terms of the distinction made above of the general versus the specific nature of the various dependent variables. Specific, achievement-related variables such as graduate grade point average and the number of times the general exams were taken are bound to be more predictable from other specific, achievement-related variables than is the more general, all-inclusive group categorization variable. Put another way,

the contention here is that although psychological variables such as motivation and perseverance play a significant role in affecting variables like grade point average, they play an even larger role in determining the ultimate outcome of one's graduate school career. Earning a Ph.D. takes at least four years of hard work, and without the motivation to stick it out, an individual's talents are not utilized.

A second factor tending to decrease the predictive power in the discriminant analysis is the simple fact that degree success is farther in time from the measurement of the independent variables than is the case with the dependent variables in the multiple regressions. But even though there is a slight loss in predictability in the discriminant group case, this is offset by the fact that a better measure of success is being obtained. Overall, it is more desirable to have a good prediction of a good criterion variable (degree success) than a better prediction of a poorer criterion (GPA).

SUMMARY AND RECOMMENDATIONS:

This paper has employed the multivariate statistical technique of discriminant analysis in an effort to uncover the predictors of a successful graduate school career. The sample of students was divided into dropout, master's, and Ph.D. groups. The variables chosen as discriminating best among the groups were, in order of selection, the Astin selectivity index, letters of reference, GRE verbal, GRE quantitative, undergraduate major, and verbal GRE-foreign student

interaction. Two discriminant functions were obtained from the analysis. The first and more important was interpreted as distinguishing the Ph.D. group from the other two groups. Each of the six independent variables was of roughly equal importance on this dimension. The second function was best thought of as a discriminator between the master's and the dropout groups. The Astin index was the primary determinant of this dimension, with GRE verbal roughly half as important. The strength of the relationship between the group criterion and the independent variables corresponded to a multiple correlation of .506. This was less than was obtained when predicting say, graduate GPA, but the closer correspondence to the concept of graduate school success was thought to offset the loss in predictability.

The study has been moderately successful, given the small sample upon which the analysis was based. But regardless of the strength of the present results, the method seems sound, and the following recommendations appear warranted:

1. Since GRE's are not required from MCS and MSTCS applicants, process these applications separately. For all other applicants, require and record data on all variables. There are enough good candidates applying that it shouldn't be necessary to admit anyone unwilling to take the GRE's or secure letters of reference. Accurate and uniform classification cannot be accomplished if data is missing. With regard to GRE scores, record the scaled scores as well as the corresponding percentiles. The scaled scores are comparable over time whereas

the percentiles technically are not, since they are based on that particular test administration only.

2. Use the classification functions of the discriminant analysis as an assessment tool in judging the applicants. Scores on these functions indicate the degree of similarity between an individual and the three group centroids. Individuals scoring highest on the Ph.D.-group function and next highest on the master's-group function are reasonably good bets to be good students. Although it might seem desirable to obtain a single predicted score for each applicant on a scale of, say, one to five, this would involve transforming the group criterion into a single continuous, interval-scale variable, which is statistically inappropriate.

The only questionable aspect in using the classification functions is that they tend to favor foreign students. The high percentage of foreign students in the Ph.D. group is manifested in the effects of the GRE verbal and verbal-foreign interaction variables. Elimination of the foreign student bias from the functions could be achieved by either computing new functions which exclude the GRE verbal and verbal-foreign variables, or by performing separate analyses for foreign and non-foreign groups, as discussed below.

3. Replicate the study at a future date. Two replications are called for. The first is a cross-validation of the results presented here. This could be done in a year or two when the entering classes of 1976 and 1977 have progressed to the point where their discriminant

group membership has become known. Cross-validation involves applying the original predictive equations to a new sample and comparing the shrinkage in predictive power to that expected by chance. Since the equations were optimized for the original sample, a loss in prediction will obviously occur.

The second and more important replication would involve merging the two samples into a single sample of reasonable size and redoing the entire analysis. The major shortcoming of this study was the small sample size upon which it was based. The results, although reasonably convincing, are to some extent the result of sampling error. Doubling or tripling the sample size would allow one to place far greater confidence in the reliability of the various statistics and coefficients. With the passage of time, it would also be possible to use the actual achievement of a Ph.D. as the basis for inclusion in the Ph.D. group.

If the sample size were sufficiently large, an additional recommendation would be to conduct separate analyses for foreign and domestic students. To the extent that different psychological factors exist within the two groups with respect to attitudes toward graduate school, separate analyses would tend to increase predictive power within each group by minimizing the effects of unmeasured, psychological variables.

References

Astin, Alexander W., Predicting Academic Performance in College, The Free Press, New York, 1971.

Nie, Hull, Jenkins, Steinbrenner, and Brent, SPSS, Statistical Package for the Social Sciences, 2nd Edition, McGraw-Hill, Inc., New York, 1975.

Stewart, D. and Love, W., A general canonical index, Psychological Bulletin, 1968, 70, 160-163.

Tatsuoka, Maurice M., Multivariate Analysis: Techniques for Education and Psychological Research, John Wiley & Sons, Inc., New York, 1971.

CASE-N	LETTERS	UGRADGPA	MAJOR	GREYERB	GREQUAN	CSGPA	CSCOURS	SEM1GPA	CUMGPA	EXAMCUM	ASTIN2
1	3	4	1	12	35	4	0	2	2	16	0
2	2	4	1	14	39	4	6	5	3	13	4
3	1	3	1	30	39	4	3	4	4	13	3
4	2	4	1	32	36	5	2	5	4	16	4
5	3	4	1	71	36	5	1	5	2	16	5
6	2	4	1	85	38	4	4	2	3	12	3
7	2	4	1	51	38	5	7	3	2	8	4
8	2	4	1	83	39	5	6	4	3	6	4
9	3	4	1	63	37	5	1	3	4	8	4
10	3	4	1	81	35	5	5	2	4	16	2
11	3	4	1	83	37	4	7	4	2	16	4
12	3	4	1	83	37	4	6	3	2	8	4
13	3	4	1	81	37	4	5	2	2	16	2
14	3	4	1	81	37	4	7	4	2	16	4
15	3	4	1	81	37	4	6	3	2	8	4
16	3	4	1	81	37	4	5	2	2	16	2
17	3	4	1	81	37	4	7	4	2	16	4
18	3	4	1	81	37	4	6	3	2	8	4
19	3	4	1	81	37	4	5	2	2	16	2
20	3	4	1	81	37	4	7	4	2	16	4
21	3	4	1	81	37	4	6	3	2	8	4
22	3	4	1	81	37	4	5	2	2	16	2
23	3	4	1	81	37	4	7	4	2	16	4
24	3	4	1	81	37	4	6	3	2	8	4
25	3	4	1	81	37	4	5	2	2	16	2
26	3	4	1	81	37	4	7	4	2	16	4
27	3	4	1	81	37	4	6	3	2	8	4
28	3	4	1	81	37	4	5	2	2	16	2
29	3	4	1	81	37	4	7	4	2	16	4
30	3	4	1	81	37	4	6	3	2	8	4
31	3	4	1	81	37	4	5	2	2	16	2
32	3	4	1	81	37	4	7	4	2	16	4
33	3	4	1	81	37	4	6	3	2	8	4
34	3	4	1	81	37	4	5	2	2	16	2
35	3	4	1	81	37	4	7	4	2	16	4
36	3	4	1	81	37	4	6	3	2	8	4
37	3	4	1	81	37	4	5	2	2	16	2
38	3	4	1	81	37	4	7	4	2	16	4
39	3	4	1	81	37	4	6	3	2	8	4
40	3	4	1	81	37	4	5	2	2	16	2
41	3	4	1	81	37	4	7	4	2	16	4
42	3	4	1	81	37	4	6	3	2	8	4
43	3	4	1	81	37	4	5	2	2	16	2
44	3	4	1	81	37	4	7	4	2	16	4
45	3	4	1	81	37	4	6	3	2	8	4
46	3	4	1	81	37	4	5	2	2	16	2
47	3	4	1	81	37	4	7	4	2	16	4

CASE-N	LETTERS	UGRADGPA	MAJOR	GREYERB	GREQUAN	CSGPA	CSCOURS	SEMIGPA	CUMGPA	EXAMCUM	ASTIM2
48	2.67	4.17	2.	93.	91.	4.14	1.	4.40	4.10	9.	7.00
49	3.33	4.71	1.	67.	99.	5.00	2.	5.00	4.90	9.	7.00
50	3.67	4.95	0.	67.	94.	5.00	1.	5.00	4.79	5.	5.00
51	2.67	4.65	1.	40.	98.	4.78	1.	5.00	5.00	5.	5.37
52	2.00	4.33	1.	12.	95.	5.00	3.	4.75	5.00	10.	5.45
53	3.00	4.75	1.	98.	99.	5.00	10.	5.00	4.93	7.	5.06
54	3.33	5.00	1.	71.	99.	5.00	3.	5.00	5.00	5.	5.36
55	2.94	4.43	1.	19.	97.	4.71	7.	5.00	4.82	5.	5.16
56	2.67	4.52	1.	17.	99.	4.67	8.	4.67	4.88	4.	5.99
57	3.00	4.65	1.	17.	91.	4.40	5.	5.00	4.85	6.	5.67
58	3.33	4.39	1.	14.	93.	5.00	5.	5.75	4.67	7.	5.00
59	3.94	4.59	1.	77.	98.	5.00	0.	5.00	5.00	8.	5.00
60	3.67	5.27	0.	84.	99.	4.71	5.	5.00	4.81	6.	5.00
61	3.33	4.39	1.	78.	98.	5.00	4.	5.00	5.00	8.	5.00
62	3.67	4.39	1.	81.	97.	4.38	5.	5.00	4.94	5.	5.00
63	2.67	5.00	0.	58.	99.	5.00	7.	5.00	4.84	4.	5.00
64	3.33	4.37	0.	77.	96.	5.00	3.	5.00	4.82	6.	5.00
65	3.33	4.60	0.	77.	94.	4.67	7.	5.00	4.70	4.	5.00
66	3.00	4.60	0.	75.	98.	5.00	6.	4.50	4.85	8.	5.00
67	3.00	4.63	0.	67.	97.	5.00	2.	4.50	4.42	4.	5.00
68	3.67	4.66	0.	83.	99.	5.00	6.	4.50	4.73	6.	5.00
69	3.00										
70	3.00										
71	3.00										
72	3.67										
73											

Astin Index Rankings of Undergraduate Colleges Attended
by the Study's Sample of Students

<u>Ranking</u>	<u>College</u>
7	Colgate
7	Princeton
7	Calif. Inst. of Tech.
7	Univ. of Rochester
7	Davidson
7	Harvard
7	Univ. of Washington
7	MIT
7	Trinity College (Conn.)
6	Whitman College (Wash.)
6	Clarkson College of Tech. (N.Y.)
6	Univ. of Michigan
6	Univ. of California - Berkeley
6	Carnegie-Mellon
6	Northwestern
6	St. Olaf College (Minn.)
5	Augustana (Ill.)
5	Univ. of Texas, Austin
5	Villanova
5	Univ. of Illinois, Urbana
5	Queens College (N.Y.)
5	Northeastern Univ. (Mass.)
5	Rockford College
5	Univ. of Connecticut
5	Carroll College (Wisc.)
5	Univ. of Kansas
4	Western Illinois Univ.
4	Baylor
4	Loyola Univ. (Calif.)
4	Univ. of Missouri, Columbia
4	Marietta College (Ohio)
4	Millikin
4	Indiana
4	Birmingham Southern College (Ala.)
3	Southern Illinois Univ.
3	Cleveland State
3	DePaul
3	Univ. of New Mexico
3	Eastern Illinois Univ.
2	Univ. of Illinois, Chicago Circle
2	Univ. of South Carolina, Lexington
2	N. Y. Inst. of Tech.
2	Univ. of Missouri, Rolla

Correlation Matrix of Independent Variables

Coefficient
(# of cases)

	Letters of ref.	Astin index	Large deg.	Undergrad GPA	Major	GRE verbal	GRE quant.	CS - GPA	# CS courses	Verbal-foreign interaction
Letters of ref.	1.0000									
Astin index	.0207 (55)	1.0000								
Large deg.	-.3444 (70)	-.1749 (57)	1.0000							
Undergrad GPA	.3887 (70)	-.1686 (57)	-.2584 (73)	1.0000						
Major	-.3065 (73)	-.0012 (73)	.2232 (73)	-.3773 (73)	1.0000					
GRE verbal	.1511 (70)	.0118 (57)	-.1609 (73)	.1835 (73)	-.1274 (73)	1.0000				
GRE quant.	.1310 (70)	.3651 (57)	.0646 (73)	.2121 (73)	-.1504 (73)	.1790 (73)	1.0000			
CS - GPA	.0416 (65)	-.0060 (55)	-.0711 (68)	.4429 (68)	-.1692 (73)	.0476 (68)	-.0533 (68)	1.0000		
# CS courses	.2168 (70)	-.0771 (57)	-.0150 (73)	.2419 (73)	-.5158 (73)	-.0419 (73)	.2442 (73)	-.0166 (68)	1.0000	
Verbal-foreign interaction	-.0382 (73)	-.1227 (73)	.1873 (73)	.0532 (73)	.1768 (73)	-.3637 (73)	-.0744 (73)	.0484 (73)	-.0794 (73)	1.0000

Correlation matrix between dependent
and independent variables
coefficient (# of cases)

	1st sem. grad GPA	cumulative grad GPA	general exam total
1st sem. grad GPA	1.0000		
cum. grad GPA	.9200 (73)	1.0000	
gen. exam total	-.6623 (73)	-.7305 (73)	1.0000
letters of ref.	.1169 (70)	.1689 (70)	-.2176 (70)
Astin index	.4103 (57)	.3582 (57)	-.2538 (57)
largest degr.	-.0196 (73)	-.0792 (73)	.1120 (73)
undergrad GPA	.2771 (73)	.3051 (73)	-.3008 (73)
major	-.1615 (73)	-.2201 (73)	.3351 (73)
GRE verbal	.0940 (73)	.0974 (73)	-.0581 (73)
GRE quant.	.3551 (73)	.3689 (73)	-.4470 (73)
CS course GPA	.1146 (68)	.0518 (68)	.0460 (68)
# CS courses	.1930 (73)	.1783 (73)	-.4370 (73)
verbal-foreign interaction	.0767 (73)	.0673 (73)	-.0517 (73)

APPENDIX 5

Standard Deviations of Independent Variables

Variables used	Group 1 Dropout	Group 2 Master's	Group 3 Ph.D.	Total
Letters of reference	.4686	.5990	.5127	.5522
Major	.5505	.7228	.5069	.6248
GRE verbal	27.7591	24.9377	33.3662	29.2268
GRE quantitative	13.7840	10.8729	2.6019	10.6318
Astin index	1.4045	1.6333	.6858	1.4270
Verbal-foreign interaction	9.0060	5.9330	16.8226	11.4210
Variables not used				
Highest prior degree	.4767	.2623	.2881	.3602
Undergraduate GPA	.4475	.4240	.2930	.4010
CS course GPA	.3367	.4565	.3621	.3913
# of CS courses	2.8410	2.0791	2.5218	2.5058

APPENDIX 6

Regression coefficients

Dependent variable = 1st semester graduate GPA

Variables in eq.	Unstandardized regression coefficients			
	1	2	3	4
Constant	3.0985	-.0295	-.5857	-.5484
Astin index	.2544	.2762	.2943	.2986
undergrad GPA		.6801	.7728	.7544
largest degree			.3676	.3249
verbal-foreign				.0068

Variables in eq.	Standardized regression coefficients			
	1	2	3	4
Astin index	.4419	.4798	.5112	.5187
undergrad GPA		.3318	.3771	.3681
largest degree			.1612	.1425
verbal-foreign				.0940

Dependent variable = cumulative graduate GPA

Variables in eq.	Unstandardized regression coefficients			
	1	2	3	4
Constant	3.3721	.3316	.3043	.8218
Astin index	.2075	.2287	.2351	.2336
undergrad GPA		.6610	.6533	.5594
verbal-foreign			.0068	.0084
undergrad major				-.1490

Variables in eq.	Standardized regression coefficients			
	1	2	3	4
Astin index	.3966	.4371	.4494	.4465
undergrad GPA		.3549	.3508	.3004
verbal-foreign			.1038	.1282
undergrad major				-.1247

Dependent variable = general exam cumulative total

Variables in eq.	Unstandardized regression coefficients			
	1	2	3	4
Constant	22.1306	21.7158	20.5155	27.5594
GRE quantitative	-.1503	-.1217	-.0761	-.0499
# CS courses		-.4974	-.5319	-.4783
Astin index			-.5770	-.7540
undergrad GPA				-1.9846

Variables in eq.	Standardized regression coefficients			
	1	2	3	4
GRE quantitative	-.4470	-.3618	-.2262	-.1484
# CS courses		-.3487	-.3728	-.3352
Astin index			-.2303	-.3010
undergrad GPA				-.2225

Data on the 5 incorrectly classified Ph.D. students

Case	Letters of ref.	Astin	Major	GRE verbal	GRE quant.	Verbal-foreign interaction
1	3.67	7.00	2	77	98	0
2	3.00	5.35	0	98	99	0
3	2.67	5.00	0	78	99	0
4	2.00	5.45	2	02	95	2
5	3.33	5.00	1	89	97	0

APPENDIX 7

Case	Func. 1 score	Func. 2 score	Grp. with highest probability	Grp. with next highest probability
1	-.595	1.892	2 - .578	3 - .374
2	-.198	.201	2 - .377	3 - .361
3	-.154	-.493	1 - .412	3 - .339
4	-.040	-.965	1 - .544	3 - .274
5	.255	.524	2 - .534	1 - .260

1. Report No. UIUCDCS-R-78-925	2.	3. Recipient's Accession No.
4. Title and Subtitle PREDICTING GRADUATE SCHOOL SUCCESS		5. Report Date June 1978
6.		7. Performing Organization Repr. No.
8. Author(s) Richard N. Nelson		9. Project/Task/Work Unit No.
10. Performing Organization Name and Address Department of Computer Science University of Illinois at Urbana-Champaign Urbana, Illinois 61801		11. Contract/Grant No.
12. Sponsoring Organization Name and Address Department of Computer Science University of Illinois at Urbana-Champaign Urbana, Illinois 61801		13. Type of Report & Period Covered MCS Report
14.		

Supplementary Notes

Abstracts

An attempt was made to identify the factors affecting graduate school success in computer science. The six most important stepwise-selected variables were the Astin selectivity index of undergraduate colleges, letters of reference, GRE verbal, GRE quantitative, undergraduate major, and verbal GRE-foreign student interaction. There are two dimensions along which the Ph.D., master's and dropout groups differed. The first separated the Ph.D. from the other two groups. All variables were of equal importance on this dimension. The second separated the master's from the dropout group. The Astin index was its primary determinant. The redundancy coefficient of .504 was less than the multiple correlation obtained when GPA was the dependent variable.

Key Words and Document Analysis. 17a. Descriptors

computer science
graduate school admissions

Identifiers/Open-Ended Terms

COSATI Field/Group

Availability Statement

19. Security Class (This Report)
UNCLASSIFIED

21. No. of Pages
38

20. Security Class (This Page)
UNCLASSIFIED

22. Price

UNIVERSITY OF ILLINOIS-URBANA



3 0112 054276586